

尺度評定と評定の確信度との関係： 質問セットによる相違のシミュレーション

妻 藤 真 彦

美作大学・美作大学短期大学部紀要（通巻第53号抜刷）

論 文

尺度評定と評定の確信度との関係： 質問セットによる相違のシミュレーション

Functional relations between scale ratings and confidence values for the ratings: A simulation study.

妻 藤 真 彦

キーワード：尺度評定、確信度、評定過程、質問紙、心理測定

尺度評定は非常に広く使われている心理学的測定であり、またこれまでその性質について詳細な検討が行われてきた (e.g., Dawson & Brinker, 1971; Dawson & Mirando, 1976; 織田, 1978; Parducci, 1965; Parducci, 1982; Petrov & Anderson, 2005; 脇田, 2004; Wedell & Parducci, 1988; for review, 妻藤, 2006)。しかし、これらの研究はすべて尺度評定過程を一種の測定装置のように扱うものであり、そこに“どのようなタイプの情報”が入力されるのかという発想はなかった (妻藤, 2007b)。

実に様々な題材について使用されているが、そのすべてが同質の情報を評定したものなのかどうかは、相当疑問だとせざるを得ない。例えば、自分自身に関する同じ内容の評定であっても、エピソード記憶に基づいて評定を行う場合と、今現在感じている意識内容を評定する場合では、評定結果が相当異なる解釈を必要とする可能性も否定できない (妻藤, 2007c)。

もっとも極端なケースとして、ある質問が例えば5点尺度での評定を要求するものであるにも拘わらず、実際に可能な判断が2値 (はい、いいえ) であるような場合、それでも5点の尺度評定が行われるなら、その評定値はその2値判断に対する確信の程度を変換して、評定値を答えるかもしれない。本稿の目的は、そのようなことが起こりやすい質問セットを見分ける方法の検討である。

他者の行動が文章で呈示されて、それがどの程度

意図的であったかを評定する課題について考えてみよう。Malle & Knobe (1997) は、このような質問のセットに対して、多数の回答者が相当決まった評定の傾向を持っているため、項目ごとの個人間平均は、項目間で有意差を示すことを確かめている。しかしこのような簡単な文章から登場人物の意図性の程度を推定できているのかどうかという問題は検討されていない。ある行動、例えば“Aさんは庭に水を撒いた”というような文であれば、意図性の程度というより“まず意図的だっただろう”と考え、また“Bさんはスーパーに車で行く途中に速度違反をしてしまった”なら、意識しないうちにスピードが出すぎていた可能性もあるので“意図的であった”か“意図的でなかった”という2値判断両方の重ね合わせ (迷い) になる可能性がある。そうだとすると、評定値が5点尺度の3点になったとき、それは、2値のどちらとも決定できないということを表すのであって、意図性が中程度だったと判断したわけではない。この場合の評定値3点は“確信がない”ことを意味しているのである (Saito, 1998; 妻藤, 2004 参照)。他方、“この演奏を退屈と感じる程度”の評定が3点であれば、“どちらともいえない”のではなく“中程度に退屈”なのかもしれない。

上記は各尺度点に付加されているカテゴリー名の問題を議論しているのではない。例えば尺度点の両端が正負の関係になっており、中央の3がゼロを表す場合を考えてみよう。例えば、“賛成”、“やや賛成”、“ど

ちらともいえない”、“やや反対”、“反対”であるとき、中央の3は、“賛成・反対が拮抗して決められない”ことを示すのか、賛成と反対の間が連続的な強さになっていて、3はその中間であるゼロ（賛成でも反対でもない：どうでもよい）のか不明なのである。もし前者であるなら“やや”が付加されているカテゴリー（4あるいは2）は、“賛成するべきだ”（“反対するべきだ”）という判断に対する確信の強さを表現したものであり、後者であるのなら、“どの程度強く賛成”（“反対”）であるかを示すだろう。

1. 評定値確信度変換モデル

妻藤（2007b）は Malle & Knobe（1997）のような行動記述文への意図性評定と、その評定に対する確信度の両方に回答する質問紙によって、両者の関係を検討した。このとき2値判断しかできず評定値は確信度が変換されて回答されているというモデル（評定値確信度変換モデル）、評定を行うときに両極は容易であるが中央付近は難しくこれが確信度に反映されるという可能性（評定困難度モデル）と、確信度は各質問項目の独自性のみによって決まるという可能性（独立モデル）を比較しようとした。このために、まず確信度変換モデルを示唆する指標を次のように工夫した。

評定も確信度も5点尺度だとすると、2値判断であるから評定値の5と1はもっとも確信度が大きく、評定値が中央の3に近づくにつれて確信度は小さくなるはずである。これを確認するには、評定値から3を減じて絶対値をとれば、確信度と1に近い相関を示すはずである（以下では、これをV字相関とする）。

ただし、評定困難度モデルでも同様のV字相関が予想されるが、このモデルの場合は評定値から確信度への影響だけではなく、各質問項目の独自性によって生じる項目ごとの確信度の相違も影響する。このように確信度は2つの独立要因の両方が働いて決定されるため、V字相関は確信度変換モデルのときよりも小さい値になるはずである。例えば、回答者自身がよく経験する行動とそうでないものでは、回答に対する確信の度合いは異なるはずである。他方、確信度変換モ

デルでは、確信度が評定値を決定するため、その質問セットの中に後述する項目間特性関係の問題がなければ、V字相関が理論上は1になるはずなのである

単純に考えると、回答者ごとに質問項目に渡るV字相関係数を出せば良いようにも思われるが、しかしこれは指標として不適当である。理論上の相関係数が1であって、それに確率変動が付加されるため、実測値が1に近い値になるという予測であるが、相関係数には1より大きな値が存在しないので、理論的に1であるなら、確率変動は係数を小さくする方向にのみ働くことになる。そして回答者ごとに計算する場合、個々の評定値が特定の質問に対して一回だけ回答された結果であるため、様々な確率変動がそのまま付加されたものから係数を出すことになる。そのため回答者ごとの係数を平均すると、実際には確信度変換モデルが正しい場合であっても、得られる平均係数はかなり1から離れた値になってしまう。

したがって確率変動の影響を何らかの形で丸めた値を用いて相関を計算する必要がある。一つの方法は各項目の個人間平均値をとることによって、上記のような評定そのものに関与する確率変動を丸めておき、評定平均と確信度平均の間の項目に渡る相関を計算することである。ただし、この場合は評定と確信の両方について項目間に平均値の有意差が必要である。

この指標は一つの候補ではあるが、単純に項目ごとの平均をとった場合、他の要因の関係で係数が大きく変動してしまうことがある。例えば、ある質問項目の特性として多くの人が大きな（小さな）評定値を答えやすく、かつ確信度が高くなる（低くなる）傾向があり、そのような項目がある程度以上含まれていると、上記のどのモデルが正しくても、それとは無関係に平均のV字相関が大きく（小さく）になってしまうからである。もちろん、この傾向は多かれ少なかれ存在するはずであり、それを上回る傾向が見出せるかどうかを検討する方法も考えられる。しかし、妻藤（2007b）は、質問セットによっては、後で述べるもう一つの指標と極端に矛盾するV字相関を示す場合があることを確認しており、この指標で議論するためには、その

質問セットに含まれる項目間の特性関係から生じる V 字相関係数の基準値が、別の手段によって分かっている必要がある。これは個人ごとに項目間の V 字相関をとる場合でも同様である。

そこで、各項目の中で評定値と確信度の関係を決定してしまい、その関係が項目間でどのように変化するかを調べる必要がある。そのために、妻藤（2007b）は項目内個人間相関係数と項目ごとの評定値平均の関係（項目に渡る相関）を検討した。この項目内個人間相関は、V 字相関ではなく、単に各尺度そのままの相関係数である。もし確信度と評定値の間に V 字相関があるなら、この相関係数は以下のように、各項目の評定値平均との間に一定の関係を示すので、その関係を表す指標を工夫すればよい。

確信度変換モデルが正しければ、評定平均が 1 に近い場合、項目内個人間相関がマイナスの値となり、この絶対値は評定平均が 3 に近づくにつれて 0 に近づき、3 を超えると正の値になって 5 の近くまで増加するはずである。評定平均が小さい項目では、個人間のバラつきが 3 より小さい値を中心とするため、スケールの全体に渡って V 字相関があるなら、評定値がこの範囲にあるデータだけを取り出せばマイナスの相関関係になる。他方評定平均が大きな項目では V 字の右側でバラつく頻度が大きいため、正の相関を示すことになる。そして平均が 3 の付近では、バラつきに V 字の左右が両方含まれるため、相関は 0 に近い値となる。（詳細は妻藤、2007b を参照されたい）。ということとは、評定平均と項目内個人間相関係数の間に強い正の相関関係（というより一次関数の関係）が得られるはずである（以下では、この相関係数をメタ V 字相関と呼ぶ）。評定困難度モデルが正しい場合は、前述の平均相関についての議論と同様の理由で、確信度変換モデルよりも小さな、ただし独立モデルよりは大きなメタ V 字相関係数が得られるはずである。

妻藤（2007b）では、登場人物の内的理由あるいは状況に基づく原因を付加した行動項目の場合、平均間 V 字相関とメタ V 字相関は、どちらも比較的大きな値を示したが、理由や原因を付加しない行動のみの記

述セットでは、平均 V 字相関が 0.34 であったのに対し、メタ V 字相関は 0.97 であった。

このメタ V 字相関は極端に大きく、確信度変換モデルを示唆する。しかし、評定困難度モデルの係数と区別するためのカットオフを設定することができていない。この 0.97 は極端に大きいとしても、理由や原因条件での値は 0.79 と 0.91 であり、これらについては、どちらのモデルを強く示唆するのも曖昧である。

本稿は、この問題を解決する前段階として、平均 V 字相関とメタ V 字相関の量的関係をシミュレーションによって検討する。確信度変換モデルが正しい場合、つまり評定値が確信度を変換した値であって、事実上同じものであるにも拘わらず、平均 V 字相関がどの程度まで小さい値になる可能性があるのか、そしてメタ V 字相関は確信度について、異なる確率分布を持つ項目群に関して、どの程度変動するかの見当をつけるためである。

2. シミュレーション

基本方針 評定値も確信度もどのような分布を示すのかは質問セットの内容によるはずであり、またシミュレーションの目的が数値的予測ではないため、簡略に分布型の相違を作り出すために 1 から 10 の間の整数乱数を生成し、その 10 個の整数の内、例えば 6 個を確信度 5 に割り当て、残りを他の確信度値に割り振るなどの方法を使用した。そのため分布はかならずしも滑らかではなく、ピーク以外の 2 つの確信度が同じ確率になる場合もある。

また、確信度変換モデルと独立モデルを比較するが、後者の場合であっても、大半の人が意図性が強い（弱い）と判断するような項目は、確信度も大きく（小さく）なる傾向があると想定するのが自然である。つまり、個々の評定においては意図性の評定値が確信度から決定されるのではなくても、大半の人が意図ありと判断するような行動は社会的経験を通じて確信度も高くなる傾向があると考えられる。そのためどちらのモデルについても、“意図あり”判断の確率が大きな項目と最小の項目で確信度の期待値が大きく、中程度の

項目で最小になるように設定された。確率変動の関係で、このようにしても平均 V 字相関は 1 にはならず、質問セットのタイプによって、かなり大きくバラつく。このバラつきがどの程度あるか、そしてそれらに対してメタ V 字相関の値が、どのくらいの大きさである、かつどの程度の安定を示すかを調べることが目的である。このようにされたので、独立モデルと言っても、平均の上では V 字相関を持つので、以下では“準独立モデル”とする。

以上のように設定されたので、準独立モデルと確信度変換モデルの比較は、後者にとって不利な比較である。ということは、今回の仮説にとって厳しい検討であるから、それでも独立モデルと変換モデルの間に明確な指標の違いが見出されたら、両者を区別する良い指標として強く推せることになる。

確信度変換モデルにおける仮想データの生成 シミュレーションは実際の実験と同様に 5 点尺度で計算された。このとき、確信度変換モデルでは、まず“意図あり”か“意図なし”の判定が起こり、その後で確信度が決定され、さらにそれが評定の 5 点尺度に変換される。

一つの質問セットに 12 個の質問項目があるものとして、まず“意図”の判定確率の期待値が、0.25 から 0.8 まで、0.05 ずつ異なるように設定された。仮想回答者数は 200 人であり、これらの期待値を生成する手順で 1 か 0 の 120 個の数値が生成された（この仮想データの実際の 1 出現頻度は、使用される乱数列によって変動する）。このように 12 個の項目が定義されるが、異なる質問セット（異なる確信度分布タイプを持つセット）ごとに異なる乱数列を用いて生成された。

異なる質問セットが次のように生成された。ピーク 5 条件では、確信度が 5 のときに最大の出現確率になり、より小さい値になるほど確率が減少するような分布生成法が用いられた。このときピークの高さが 3 通り設定され、これによって確信度の期待値が 3 通りでき、それらの各々を用いて 4 つの項目が各々異なる乱数列によって生成されて 12 個となった。同様の方法を用いてピークが 3 になるセットも作られたが、この

セットではピークが高い項目は左右対称、小さい項目では右または左に 2 番目に大きい確率が片寄るようにして、期待値の異なる項目が 3 通り作られた。またピーク 5 とピーク 3 の実現値（各仮想回答者の値）から 0.5 の確率でランダムに選んで、中間分布が作られた。さらに、ピーク 5 のセットと 3 のセットから 0.5 の確率でランダムに項目を選び、混合セットが作られた。このように 4 種の異なる特性を持つ質問セット（条件）が作られ、かつ異なる乱数列に基づいて、4 条件すべてが 2 セット作られ、各々シミュレーション 1 と 2 として使用された。

以上の各質問セットについて、ある仮想回答者のある項目に対する意図判定が 1（意図的だという 2 値判断）である場合、確信度の 5 と 4 が意図性評定値の 5（まったく意図的）となり、3 と 2 が評定値の 4、確信度 3 が評定値の 3 になった。意図判定が 0（意図的ではないという判断）のときは、確信を持って意図的でないと思うときに、意図性評定値は最低になる。したがって、確信度の 5 と 4 が意図性評定値の 1 に対応し、確信度の 1 が意図性評定値の 3 に変換される。つまり、このモデルでは“意図的だ”という判断に対する“確信なし”と、“意図なし”に対する“確信なし”は、意図性の程度評定をさせると同じ値（評定値 3）になると仮定している。このように、意図あり判断の確信 1 と意図なし判断の確信 1 の 2 つの場合が意図性評定値の 3 に変換されるので、全ての意図性評定値は、確信度 2 つ分に対応することになる。

準独立モデルにおける仮想データの生成 このモデルは、変換モデルと比較するために、確信度は同じものが使用された。単に同じ生成法を使用するのではなく、まったく同じ数値が使用された。そして評定値は、変換モデルと同じ意図判定数値（1,0）に対して、確信度と同じ生成法だが、異なる乱数列を用いて 5 点尺度が作られた。

3. シミュレーションの結果

メタ V 字相関がどのようなものになるかを示すために、確信度変換モデルの、各項目の評定平均と項目

内個人間相関係数の関係を表すシミュレーション結果の一部を図1に示す。

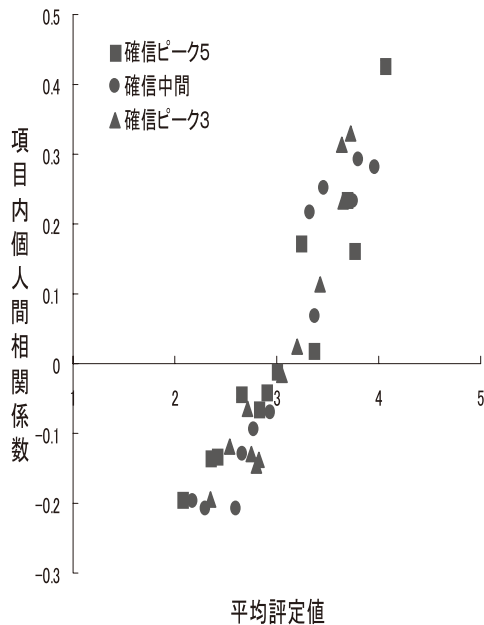


図1 シミュレーション1：項目の評定平均に対する
評定値・確信度の項目内相関係数

実データ（妻藤、2007b）とよく似た結果であり、メタV字相関係数も条件間で、実データと近い値でバラついている（表1）。確信度変換モデルのシミュレーション1では、ピーク3（歪度3種）の項目セットの平均V字相関とメタV字相関の両方が、実データの“行動のみ”条件と非常に近い値になっている。モデルでは、0.35と0.95であるが、実データでは0.34と0.97であった。他の条件では、全体にメタV字相関が全て0.9より大きく、ほぼ0.95の周辺になっている。他方実データでは理由条件で0.79、原因条件で0.91であり、前者はいささか小さい。また、実データの平均V字相関は“理由付き”で0.71、“原因付き”で0.86であり、各々、シミュレーション1の中間セットとピーク5セットにかなり近い値を示している。

他方、準独立モデルでは、平均V字相関が0.65か

ら0.95の範囲であるのに対して、メタV字相関はマイナス0.4から0.45の範囲に入っている。

表1. シミュレーションの結果

	確信度分布の型 ^{c)}			
	ピーク5	中間	ピーク3	混合
確信度変換モデル（シミュレーション1）				
rV ^{a)}	0.86	0.77	0.36	0.88
rmV ^{b)}	0.93	0.98	0.95	0.96
確信度変換モデル（シミュレーション2）				
rV	0.91	0.86	0.58	0.59
rmV	0.94	0.95	0.95	0.96
準独立モデル（シミュレーション1）				
rV	0.88	0.69	0.67	0.86
rmV	0.44	-0.37	0.18	0.41
準独立モデル（シミュレーション2）				
rV	0.94	0.69	0.67	0.89
rmV	0.12	-0.15	0.15	-0.24

（注）^{a)} rV は、各項目の評定値平均 - 3 と確信度平均の、項目に渡る相関係数（平均V字相関）

^{b)} rmV は、各項目内での評定値・確信度の被験者に渡る相関係数（項目内個人間相関）と、各項目の評定値平均の間の、項目に渡る相関係数（メタV字相関）

^{c)} 各々の型の中には、ピークの高さや歪度の相違（異なる期待値）を持つ項目が含まれている

4. 考 察

今回のシミュレーションが、平均V字相関に影響するような、確信度変換過程以外の要因を全て考慮できているかどうかは不明である。しかし、現在の理論探索段階の検討においては、この値を一つの目安として使うことが可能ではないと思われる。

もし確信度と評定値の相関関係に影響する要因がほぼ組み込まれているとするならば、モデルと実測の対応を検討しようとするときに、平均V字相関がほぼ同程度である条件について、メタV字相関が実測とどのくらいズレているかが、一つの見当の付け方であろう。シミュレーション1のピーク3は、実測の“行動のみ”条件と両方の値が良く対応しており、少なく

ともこの条件は、実際に確信度が評定値に変換された可能性が極めて強いと言ってよいであろう。

妻藤（2007b）は、この条件において2つの指標が極端に乖離した原因として、各質問項目が持つ特性の問題を指摘した。ピークが大きな確信度のところにある項目セットでは、そのピークになる確率が項目ごとにかなり異なっている、全体として平均V字相関があまり小さくはならない。混合セットの場合は0.6程度の値になることもあるので、ピークが大きな確信度になる項目が半分以上含まれているセットの場合には、その含まれ方によってかなり変動すると言ってよいであろう。他方、ピークが3で項目によって歪度が異なるセットの場合には、0.36あるいは0.58のように、比較的小さな値になりやすい。

おそらく“理由付き”条件や“原因付き”条件の場合は、自分の過去の行動の記憶に基づいて、そのような理由や原因が付加されておれば、その行動が意図的であったかそうでなかったかの記憶情報が得やすく、確信度が最大値か、それに近い値の分布になる可能性が大きいのだと考えられよう。

それに対して、“行動のみ”では、どのような理由・原因であるのかが明示されていないため、項目によっては理由が推測されやすかったり、原因であったりするであろうし、どちらともつかない項目もあると考えられる。すると、そのような項目セットの場合には、確信度の出現確率ピークは、全体として低い値になりやすく、項目によってはフラットに近づくかもしれない。そのため、全体としては項目ごとの確信度の期待値と評定値の期待値が、V字関係を満たすようになっているにも拘わらず、確率変動が大きく影響して平均V字相関を引き下げてしまうことがある。

実測の“理由付き”条件では平均V字相関がシミュレーション1の中間セットに近い値であるが、メタV字相関は0.79であるのに対し、シミュレーションでは0.98である。このような場合、確信度変換モデルではなく、評定困難度モデルの方をより強く示唆するという解釈も可能であろう。今回は評定困難度モデルのシミュレーションは行っていないが、理論上メタ

V字相関は準独立モデル（0.44以下）よりも大きく、確信度変換モデル（0.93以上）より小さいはずであるから、もし今回のシミュレーションのみで判断するとしたら、“理由付き”条件について可能性が高いのは、評定困難度モデルの方であると言えるだろう。

もちろん今回の結果は探索的なものであるから、これだけで、強い結論に至ることは出来ないが、今後項目セットの性質と各項目の特性について、より詳細な分析とシミュレーションの回数を増やすこと、あるいは数理モデル化することで、メタV字相関係数のカットオフ値を見出せるであろう。このようなカットオフが確定できれば、質問内容の性質によって2値判断しかできなかったり、評定の難しさが尺度点によって異なったり、そのようなことのない場合の区別が出来るようになる。このような方法を作り出せれば、その質問内容の特性をこれまでとは別の観点で分析でき、意識との関連（e.g., 妻藤, 1992; 妻藤 2007a）についても、実証的な検討の幅を広げることができるようになる。

要 約

非常に広く使われている尺度評定について、実際にはどのような情報を評定したものであるのかを検討する一つの方法として、評定値と確信度の関係が検討された。質問のセットによっては2値判断しかできず、その判断に対する確信度が変換されて評定値になる（確信度変換モデル）場合があるかどうかを示唆する2つの指標（妻藤, 2007b）について、確信度変換モデルと準独立モデルのシミュレーションが行われ、特定の質問セットについて、どのモデルが当てはまるかを示唆する指標のカットオフ値を決定する方法が提案された。ただし、今回の結果だけではまだ全ての要因が検討されたわけではなく、今後シミュレーションの改善あるいは数理モデル化が必要である。

文献表

- Dawson, W.E., & Brinker, R.P. (1971) . Validation of ratio scales of opinion by multimodality matching. *Perception & Psychophysics*, **9**, 413-417.
- Dawson, W.E., & Mirando, M.A. (1976) . Inverse scales of opinion obtained by sensory-modality matching. *Perceptual and Motor Skills*, **42**, 413-425.
- Malle, B.F. (1999) . How people explain behavior: A new theoretical framework. *Personality and Social Psychology Review*, **3**, 23-48.
- Malle, B.F., & Knobe, J. (1997) . The folk concept of intentionality. *Journal of Experimental social psychology*, **33**, 101-121.
- 織田輝準 (1978) 評定尺度による判断過程の研究 教育心理学研究、**26**, 142-151.
- Parducci, A. (1965) . Category judgment: A range-frequency model. *Psychological Review*, **72**, 407-418.
- Parducci, A. (1982) . Category ratings: Still more contextual effects! In B. Wegener (Ed) , *Social attitudes and psychophysical measurement*. Hillsdale: Lawrence Erlbaum. Pp.89-105.
- Petrov, A.A. & Anderson, J.R. (2005) . The dynamics of scaling: A memory-based anchor model of category rating and absolute identification. *Psychological Review*, **112**, 383-416.
- 妻藤真彦 (1992). 根拠を述べることができない確信と「意識様態」。美作女子大学・美作女子大学短期大学部紀要、**38**, 1-10
- Saito, M. (1998) . Fluctuations of answer and confidence rating in a general knowledge problem task: Is confidence rating a result of direct memory-relevant-output monitoring? *Japanese Psychological Research*, **40**, 92-103.
- 妻藤真彦 (2004). 確信度評定のメカニズムと理論的問題. 風間書房
- 妻藤真彦 (2006). 尺度評定過程への入力情報の問題. 美作大学・美作大学短期大学部紀、**51**, 1-10.
- 妻藤真彦 (2007a). 認知・学習過程の理論的観点から見た尺度評定と潜在的測定の関係. 美作大学・美作大学短期大学部紀要、**52**, 1-10.
- 妻藤真彦 (2007b). 質問紙評定過程における参照情報－他者行動の評定－. 心理学研究、**77**, 541-546.
- 妻藤真彦 (2007c). 記憶に基づく自己評定の確信度変動と評定値の関係. 日本心理学会第71回大会発表論文集、583.
- 脇田貴文 (2004). 評定尺度法におけるカテゴリ間の間隔について. 心理学研究、**75**, 331-338.
- Wedell, D.H., & Parducci, A. (1988) . The category effect in social judgment: Experimental ratings of happiness. *Journal of personality and Social Psychology*, **55**, 341-356.